Data Lake Naming Conventions for S3 Buckets

Definition of Terms

Source (Customer or Data Provider)

This is the top level descriptor of data within an environment and is used to designate either the "tenant" for whom the data is being managed, or will be used to designate the provider of data in the case that the data is not already associated to a specific customer.

Examples:

For a full list of customers see the Customer reference page.

Phase

Phase is used to manage the different representations of data as it is cleansed, enriched, and loaded through the pipeline.

Examples:

- raw
- cleansed
- enriched
- error

Activity Type

Activity type is used to describe the internal categorization of how the data will be used. Business rules for what to do with incoming data will be largely driven off of the activity type associated to the load.

Examples:

- encounter
- prospect
- newmovers
- dns (do not solicit)
- callcenter
- marketinglist
- deceased
- hra (health risk assessment)
- employeeroster
- eventregistration
- lead

Format

Format will be used to designate the template or basic set of attributes tracked in a given schema. If a template is used from a vendor that is partnered with a customer, the format will be the name of the vendor. If a template is provided internally by Influence Health, the format will be "influencehealth". If a template is the default from a single source that only provides only one format, it will be "standard". Formats can have multiple versions that evolve over time.

Examples:

- influencehealth (internal templates for encounter(a.k.a utilization), employee roster, marketing list, donor list)
- standard (experian's prospects, or social security administration's death-master-file)
- beryl (callcenter)
- foundation (dns)
- medicom (hra)

Version

Version will be used to explicitly track the changes to a format over time in a way that allows for old schema versions and newer schema versions to be loaded.

Schema

Schema represents a specific combination of source, type, format, and version. A schema is the most granular definition of a file and is used by cleanse jobs to map incoming attributes to an Activity.

Batch

Batch is used to specify a grouping of data that is loaded together as an atomic unit. Batches can be tracked in either monthly or daily increments that should be configured based on the type and format of the data. Batches are the most granular level at which data can be loaded through the system. Reloading a batch will replace and reprocess all Activities for that batch. Person information will be updated according to the new batch data. Note: Existing updates to persons from the previous load of a batch will not be removed, but anything that causes a person to change will be overwritten by the new batch load. (i.e. removing an activity from a batch and reloading it will not undo a change that was applied when the batch was originally processed)

Examples:

- 2017-12 (monthly batch for april 2017)
- 2018-05-04 (daily batch for may 4th 2018)

Batch ID

A batch id is created at the time a batch is loaded through the Data Pipeline for a customer. This id is then used to track the activities that are part of a given load for auditing, quality control, and reporting. A batch id consists of:

`source-activitytype-format-batch`

Examples:

- chomp-encounter-influencehealth-2018-03
- northwell-dns-foundation-2018-01
- tanner-encounter-redox-2018-05-20

S3 Bucket Names

Data will be hosted on S3 using the following file name conventions and directory pattern. This pattern is natively integrated into the CDP application and is used to provide a well defined

```
s3://edh-data-<env>/<source>/<status>/<activitytype>/<format>/<batch>/...
```

Note: Data that is not specific to a customer (such as experian and social security administration data) will be stored in a bucket based on the organization that provides the data

Cleansed Parquet files

Please note that parquet files are stored as partitions, so you will see multiple pieces with a part-number-(GUID) prefix within each "file". An example would be:

```
s3://edh-staging-data/northwell/cleansed/encounter/influencehealth/2018-03
.parquet/
    _SUCCESS
    part-00000-216e2843-88aa-4058-8c93-16192355dd85.snappy.parquet
    part-00000-adbb37d7-b461-4036-92f8-5a56d760872a.snappy.parquet
    part-00000-9a453efe-021d-43b4-a384-bfc9d4a3f41b.snappy.parquet
```

Error CSV Files

When a file contains rows with errors, these rows will be saved to a specific file in the error bucket. The error file CSV should contain the batch-id of the original load so that it can be easily associated later if corrected.

```
s3://edh-<env>-data/<source>/error/<activitytype>/<format>/<batch>/<batch-id>-invalid-records.csv
```

Example Scenarios

1) A new production encounter load for northwell in march of 2018 should contain all files (visit, demog, physician, cpt, dx, etc...) inside the bucket

```
s3://edh-prod-data/northwell/raw/encounter/influencehealth/2018-03/
```

2) Once the new utilization files have been processed by the cleanse job, the resulting file will be loaded to

```
s3://edh-prod-data/northwell/cleansed/encounter/influencehealth/2018-03.parquet/
```

3) When a new file is cleansed there is also a possibility of producing error csv file containing bad records with a failure reason at

```
s3://edh-prod-data/northwell/error/encounter/influencehealth/2018-03-invalid-records.csv
```